Optical Engineering 49(3), 037401 (March 2010)

# Empirical mode decomposition-based facial pose estimation inside video sequences

**Chunmei Qing**
**Jianmin Jiang**
University of Bradford
Digital Media and Systems Research Institute
Bradford BD7 1DP, United Kingdom
qingchunmei99@hotmail.com

**Zhijing Yang**
Sun-yet Sen University
School of Mathematics and Computing Science
Guangzhou, 510275 China

**Abstract.** We describe a new pose-estimation algorithm via integration of the strength in both empirical mode decomposition (EMD) and mutual information. While mutual information is exploited to measure the similarity between facial images to estimate poses, EMD is exploited to decompose input facial images into a number of intrinsic mode function (IMF) components, which redistribute the effect of noise, expression changes, and illumination variations as such that, when the input facial image is described by the selected IMF components, all the negative effects can be minimized. Extensive experiments were carried out in comparisons to existing representative techniques, and the results show that the proposed algorithm achieves better pose-estimation performances with robustness to noise corruption, illumination variation, and facial expressions. © *2010 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.3359510]

Subject terms: empirical mode decomposition; intrinsic mode function; mutual information; facial pose estimation; feature face; bandpass filter.

Paper 090644RR received Aug. 25, 2009; revised manuscript received Jan. 14, 2010; accepted for publication Jan. 16, 2010; published online Mar. 22, 2010.

## 1 Introduction

Facial pose is an important visual cue to determine a human's identity and activities across different scenarios. Estimation of facial poses from video sequences is important for both computer vision and multimedia content analysis, such as scene understanding, event estimation, etc., or activity analysis in video surveillances.[1,2] Over the past decades, facial pose estimation remains an active research area in which a range of techniques has been reported to investigate the pose-estimation problem, such as support vector classification,[3] eigenspace from Gabor filters,[4] manifold learning,[5] independent component analysis,[6] and a two-stage framework based on Gabor wavelets, bunch graphs,[7] etc.

Recently, mutual information (MI) is used to extract facial poses from video sequences.[8] MI is widely used as a powerful tool for finding similarities between two entities. However, the reported work[8] directly applied MI to the original facial images, and as a result, its estimation rate is subject to illumination changes and noises. In order to find the fundamental nature of the facial images for accurate facial pose estimation, which is robust to illumination variations and noises, a feature extraction processing stage is necessary. Traditionally, decomposition techniques, such as Fourier decomposition or wavelet decomposition using basis functions, are selected to analyze real-world signals as powerful tools for feature extraction.[9,10] However, the main drawback of those approaches is that the basis functions are fixed and not necessarily match the varying nature of those input signals, such as facial images. In early studies, Fourier analysis has been the dominating signal analysis tool for feature extraction. However, the signal to be analyzed must be strictly periodic or stationary; otherwise, the resulting Fourier spectrum will fail to characterize the input signals and thus be inappropriate for further processing, such as pose estimation, which is often nonstationary, nonlinear, and hence the frequency components change with time. Out of all existing time-frequency analysis methods, Wavelet transform maybe the best candidate and has been widely reported for feature extraction in images and videos. However, due to the limitation of Heisenberg-Gabor inequality, wavelet transform cannot achieve fine resolutions in both time domain and frequency domain simultaneously.[11] Therefore, it fails to separate those signals with high-frequency noises, where the time scales are often too small.

To overcome this problem, we introduce a new concept of empirical mode decomposition (EMD) based on the work reported by Huang et al.[12] for nonstationary and nonlinear signal processing to extract features for a new algorithm design toward efficient and effective pose estimation. Motivated by the fact that EMD can decompose any complicated signal into a sum of intrinsic mode function (IMF) components holding the highest local frequency from the rest, we combine the strength of EMD and MI to propose a new pose-estimation algorithm by utilizing EMD as a bandpass filter in this paper. While MI is applied to characterize the similarity between poses, EMD is exploited to decompose the input video frames into IMF components, enabling the proposed algorithm to minimize the effects of noise and illumination variations when extracting the feature face for pose estimation. Compared to existing approaches, extensive experiments support that our proposed algorithm is robust to noises, illuminations, facial expressions, and significantly performs better than representative existing benchmarks.

The rest of paper is organized as follows: Section 2 de-

scribes the EMD algorithm and the details of extracting the feature faces based on EMD. Section 3 presents the algorithm for automated facial pose estimation. In Section 4, experimental results and relevant discussions are described, and finally, conclusions are given in Section 5.

## 2 Feature Face Extraction

### 2.1 *EMD*

The joint space–spatial frequency representations have received special attention in the fields of image processing, computer vision, and pattern recognition. Because there exist some crucial restrictions in Fourier transform and wavelet transform explained in Section 1, they are not appropriate for post estimation. Huang et al.[12] presented a multiresolution decomposition technique, referred to as EMD, which was originally proposed for the study of ocean waves,[12] and was later found potentially applicable to geophysical exploration, underwater acoustic signals, noise removal filter, biomedicine, pattern recognition, etc.[13–17]. The major advantage of using EMD is that the basis functions can be directly derived from the signal itself based on the local characteristic time scale of the data, which provides better characterization than those given in advance.[12] It is a fully data-driven approach and often brings not only high decomposition efficiency but also sharp frequency and time localizations.

Essentially, EMD decomposes a signal into a sum of oscillatory functions, namely, IMFs, which (*i*) have the same number of extrema and zero crossings or differ at most by one and (*ii*) are symmetric with respect to local zero mean, where the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero at any point. The IMF components are obtained from the signal by the means of an algorithm called the sifting process. This algorithm extracts each mode locally and excludes the highest-frequency oscillations out of the original signal.

Given those two definitive requirements of an IMF, the sifting process to extract the first IMF $c_1(t)$ from a given signal $x(t)$, $t=1,\ldots,T$ is designed as such that, initially, $s(t)=x(t)$, and then is subjected to an iteration process that can be described as follows:

1. Identify all the local maxima and minima of $s(t)$.
2. Generate its upper and lower envelopes, $e_{\mathrm{up}}(t)$ and $e_{\mathrm{low}}(t)$, by cubic spline interpolation.
3. Calculate the mean envelope by $m(t)=[e_{\mathrm{up}}(t)+e_{\mathrm{low}}(t)]/2$.
4. Sifting: $d(t)=s(t)-m(t)$.
5. Check the properties of $d(t)$:

   If $d(t)$ is not an IMF, then let $s(t)=d(t)$ and go back to step 1.
   If $d(t)$ is an IMF, then let $c_1(t)=d(t)$ and end the iteration.

The residue $r_1(t)=x(t)-c_1(t)$ is treated as the new input [i.e., $s(t)=r_1(t)$], and the same iteration is applied to the new input to extract the next IMF and produce the next residue. Such an iteration carries on until the sifting process is stopped by any of the following criteria: after extracting $n$ IMFs, the residue, $r_n(t)$ is either an IMF or a monotonic function. $r_n(t)$ can be interpreted as the dc component of the signal. After that, the original signal can be reconstructed via the following equation:

$$x(t)=\sum_{i=1}^{n} c_i(t)+r_n(t). \tag{1}$$

It should be noted that in step 2 we adopt the boundary conditions proposed in Ref. 18 to treat the end effects of the spline fitting, which is achieved by adding extrema by mirror symmetry with respect to the extrema that are closest to the edges. EMD aims to capture information about local trends in the signal by measuring and quantizing oscillations. Such oscillations can be quantized by a local high frequency or local detail and correspondingly a local low frequency or local trend. The source signal being decomposed into these local details and trends can be iteratively reduced to characteristic signals. If we use EMD to decompose facial images into their IMFs, then there is a strong likelihood that the effects of noise and illumination will be isolated into one or more IMFs.

In order to better understand the way EMD behaves in stochastic situations involving broadband noise, Ref. 19 reports on numerical experiments based on fractional Gaussian noise (FGN). In such a case, it turns out that EMD acts essentially as a dyadic filter bank resembling those involved in wavelet decompositions. FGN is defined as the increment process of fractional Brownian motion.[20] In discrete time, FGN corresponds to a time series $\{y_H[n],\ n=\ldots,-1,0,1,\ldots\}$ indexed by a real-valued parameter $0<H<1$ (its Hurst exponent), and such that its autocorrelation sequence $R_{x_H}[k]:=E\{y_H[n]y_H[n+k]\}$ can be worked out as

$$R_{x_H}[k]=\frac{\sigma^2}{2}(|k-1|^{2H}-2|k|^{2H}+|k+1|^{2H}). \tag{2}$$

As is well known, the special case $H=1/2$ reduces to white noise, whereas other values induce nonzero correlations, either negative when $0<H<1/2$ or positive when $1/2<H<1$ (long-range dependence).

Figure 1 (Ref. 19) illustrates the case of fractional Gaussian noise, where EMD can be interpreted as a filter bank of overlapping bandpass filters for modes of indices $k\geqslant 2$, the mode 1 corresponding essentially to a half-band high-pass filter (although it contains a non-negligible low-pass part in the lower half-band). Moreover, each mode of index $(k+1)$, $k\geqslant 2$, occupies a frequency domain that is roughly the upper half-band of that of the previous residual of index $k$. The collection of all such filters tends to organize in a filter bank structure that is reminiscent of what is classically observed in wavelet decompositions in similar situations.[21] It is worth pointing out that similar results have been obtained independently by Wu and Huang[22] in the case of white noise.

On the basis of the above observation, if we set $c_{n+1}=r_n(t)$ as the last IMF, then a general purpose time-space filter can be designed as

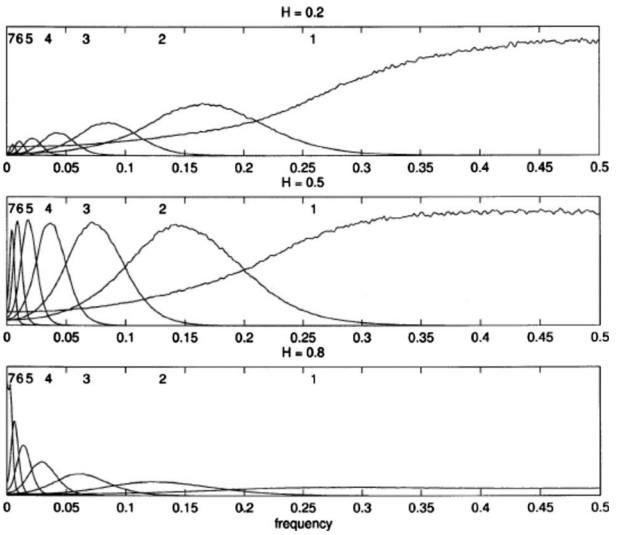


**Fig. 1** Illustration of EMD behavior, where 5000 independent time series of 512 points each have been generated, the average spectra of the seven first IMFs are plotted as a function of normalized frequency, and the value of Hurst exponent varies in 0.2, 0.5, and 0.8.

$$x_{lh}(t) = \sum_{i=l}^{h} c_i(t), \tag{3}$$

where $l,h \in \{1,\ldots,n+1\}, l \leqslant h$. For example, when $l=1$ and $h<n+1$, it is a high-pass filtered signal; when $l>1$ and $h=n+1$, it is a low-pass filtered signal; when $1<l \leqslant h<n+1$, it is a bandpass filtered signal. From Eq. (3), it can be seen that EMD allows us to selectively reconstruct the input signal by ignoring those IMFs whose contributions to the signal are undesirable. For pose estimation, such contributions include unwanted noise, illumination changes, and expression variations. In this paper, Eq. (3) forms the basis functions for representing face data as described below, where we use it as a bandpass filter.

### 2.2 *Feature Face Extraction*

In general, a facial image is often contaminated by noise and uneven illumination in many multimedia processing applications, such as computer-aided media content analysis, determination of human identity, activities across different scenarios, etc. Therefore, a feature face with noise and illumination invariant in the facial pose estimation is especially important in these applications. From Refs. 17 and 23, it can be deduced that majority of noise is represented by the highest local information (the first IMF) and the majority of illumination effect is represented by the last several IMFs, which motivated us to apply EMD as a bandpass filter to decompose input signals, and only those IMFs that describe the distinct facial pose characteristic are used as discriminating features for facial pose estimation.

To illustrate how EMD can be used as a bandpass filter, the process of decomposing signals extracted from facial images are shown in Fig. 2. According to the property of the EMD procedures, the data are decomposed into several fundamental components, each with a distinct time scale. One example is shown in Fig. 2(c), which includes the original signal extracted from the middle row of Fig. 2(a) and its decomposed four IMFs and one residue. More specifically, the first IMF associated with the smallest time scale corresponds to the fastest time variation of data. As the decomposition process proceeds, the time scale is increasing, and hence, the mean frequency of the mode is decreasing. Figure 2(d) shows us the EMD result of the middle-row signal extracted from Fig. 2(b), which has illumination changes from right to left. Comparing Figs. 2(c) and 2(d), it can be seen that the IMFs decomposed from these two different signals are nearly the same. The difference just exists in the residues, which reflect the illumination tendency of the signals such as the dc component in the data. It can be required that the inhomogeneous illumination correction can be performed after subtracting the residue from the original signal.

Because the EMD sifting process first extracts the highest frequency, the first modes correspond generally to the noise. It can be observed by comparing Fig. 2(c) with Fig. 2(f), which is the EMD result from a Gaussian white-noise corrupted signal. Figures 2(c) and 2(f) have very similar low-frequency IMFs and residues. The difference is that Fig. 2(f) has one more decomposed IMF, which is reflected by the first IMF. To illustrate what the role of the first IMF in Fig. 2(f) is, the reconstruction is shown in Fig. 2(g) in which the original data without noise from Fig. 2(c) is plotted as a blue line and partial sum of the IMFs is displayed as a red line. The red line is the summation of the second to the fifth IMFs and the residue from Fig. 2(f). It can be seen that it matches the shape of the original data well. Therefore, the information of the noise is contained in the very first modes. In fact, illumination variation can be treated as a type of noise, which is related to low-frequency oscillations or signal tendency. Therefore, in this paper we adopt $l=2$ and $h=n$ in Eq. (3) to form the bandpass filter for low-level processing (filtering or denoising) to reduce illumination and noise effects.

There are a number of 2-D EMD techniques that have been reported in the literature. For example, Nunes et al.[24] proposed a 2-D EMD using a radial basis function for surface interpretation; Damerval et al.[25] presented a fast algorithm for bidimensional EMD, Xu et al.[26] provided a 2-D EMD by finite elements; and Xu et al.[27] proposed an improved bidimensional EMD based on structural extrema. However, as reported by Huang and Wu in Ref. 28, all these 2-D approaches are computationally expensive and provide different results for the same image because of different interpolation methods. Though the spline-fitted surface serves the purpose well, the fittings offer only an approximation and could not go through all the actual data points. We have also carried out empirical studies on 2-D EMD, which indicate that there exist a range of challenges and difficulties for 2-D EMD, and hence, we adopted the original version of EMD proposed by Huang et al. in Ref. 12 to decompose images row by row. It is a direct method to extend 1-D EMD to process 2-D images, in which our essential purpose is to illustrate one specific aspect of EMD (i.e., the way it filters out the noise) in a real-world application (facial pose estimation). To associate with facial pose estimation, by treating each row of a facial image separately, we can string 2-D facial images into 1-D vectors. Application of EMD to these vectors yields a set of vector IMFs that are then reshaped into matrix IMFs, as

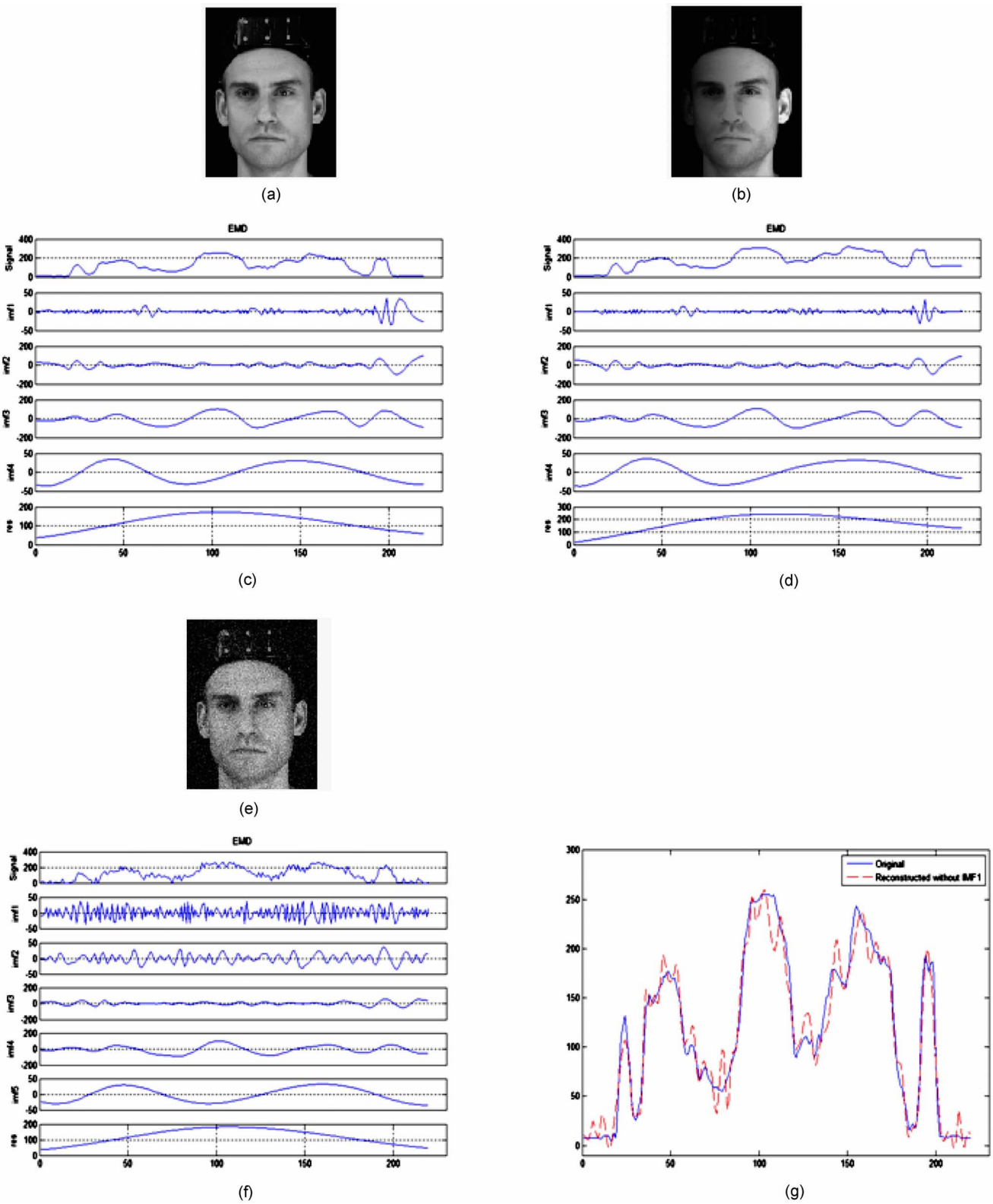


**Fig. 2** (a) The original image; (b) the original image with light from right to left; (c) EMD result of the middle row signal of (a); (d) EMD result of the middle row signal of (b); (e) the noise corrupted image; (f) EMD result of the middle row signal of (e); (g) the original signal (solid line) of (c) and the reconstructed signal without the first IMF (dashed line) of (f).

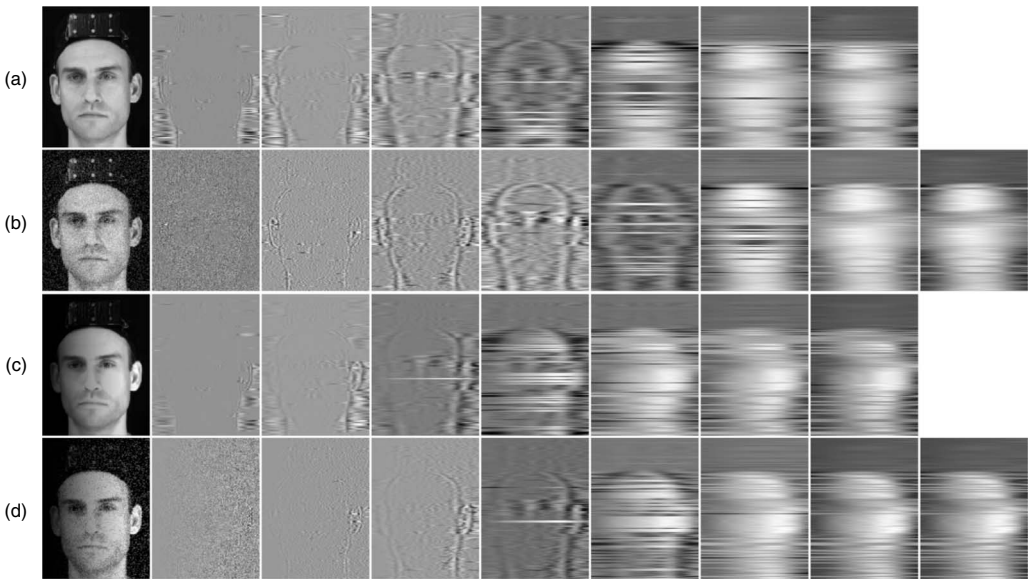


**Fig. 3** (a) The original facial image and its decomposed IMFs; (b) the noise corrupted image and its decomposed IMFs; (c) the original image with light from right to left and its decomposed IMFs; and (d) the image contaminated by noise and uneven illumination and its decomposed IMFs.

shown in Fig. 3, in which the facial face is decomposed into several IMF faces showing the details from fine to coarse.

By comparing Figs. 3(a) and 3(b), we can see that the first IMF face of the noise corrupted image contains the majority of the highest-frequency oscillations (i.e., the variation of noise). Although its second IMF face is also affected by noise oscillations, it still gives us the characteristics of this face image. The remaining IMFs of both Figs. 3(a) and 3(b) match very well. Figure 3(c) is an example to investigate the effect of illumination changes. With the comparison of the last images of Figs. 3(a) and 3(c), it can be seen that the residue is responsible for the majority of the illumination effect. Figure 3(d) illustrates the EMD results of one facial image contaminated by noise and uneven illumination at the same time. It can be seen that both types of noise can be separated well. From the above analysis, it can be concluded that the first IMF represents the effect of high-frequency noise and its residue relates to the illumination tendency. Therefore, when facial images are reconstructed without those two components, their processing could be made robust and, hence, their performances in pose estimation could be improved.

Given the facial image, $X=[x_1(t),\ldots,x_i(t),\ldots,x_M(t)]'$, $t=1,\ldots,N$, each row $x_i(t)$ can be decomposed into $x_i(t)=\sum_{j=1}^{n_i} c_{ij}(t)+r_{n_i}(t)$ by the EMD. Removing the first IMF $c_{i1}(t)$ and the residue $r_{n_i}(t)$ of each row, the feature face can be reconstructed by

$$\tilde{X}=\begin{bmatrix}\tilde{x}_1(t)\\ \cdots\\ \tilde{x}_i(t)\\ \cdots\\ \tilde{x}_M(t)\end{bmatrix}=\begin{bmatrix}\sum_{j=2}^{n_1} c_{1j}(t)\\ \cdots\\ \sum_{j=2}^{n_i} c_{ij}(t)\\ \cdots\\ \sum_{j=2}^{n_M} c_{Mj}(t)\end{bmatrix}. \quad (4)$$

Selective reconstruction of facial images using IMFs that do not contain high-frequency noise and illumination effects enables us to reconstruct the fundamental nature of the data for accurate facial pose estimation.

## 3 Facial Pose Estimation

Existing work[8] on pose estimation is established on the setup that, given $N$ people for whom the pose is to be

**Fig. 4** Six poses in the MPI database.

**Fig. 5** Examples of 38 different facial expressions in the MPI video sequence.

estimated, a database of template images: $O=O_1 \cup O_2 \cup \ldots \cup O_N = \cup_r O_r$, is created, where $O_r = \cup_j O_{j,r}$ $(j=1,\ldots,J)$, contains $J$ different poses for the $r$'th person. Therefore, let the input video for the $r$'th person be partitioned into a set of $K$ frames $F_{1r},\ldots,F_{Kr}$, his or her pose can be estimated via examination of each video frame $F_{kr}$ $(k=1,\ldots,K)$ in comparison to the images in the set $O_r = \cup_j O_{j,r}$ $(j=1,\ldots,J)$ to determine which pose set $O_{j,r}$ provides the best match between the template and the input video frame.

In Ref. 8, mutual information is used to measure the dependency of the information contained in a test frame $F_{kr}$ and a reference frame $f_i$, $f_i \in O_{j,r}$, which is defined as follows:[29,30]

$$\mathrm{MI}(F_{Kr},f_i) = H(F_{kr}) + H(f_i) - H(F_{kr},f_i), \tag{5}$$

$$H(F_{kr}) = -\sum_u p_{F_{kr}}(u)\log p_{F_{kr}}(u), \tag{6}$$

$$H(F_{kr},f_i) = -\sum_{u,v} p_{F_{kr}f_i}(u,v)\log p_{F_{kr}f_i}(u,v). \tag{7}$$

In order to calculate the MI between these two video frames, which have $L$ pixels each, they consider pixel values as outcomes $A=\{a_1,a_2,\ldots,a_L\}$ of a random variable. The probability $p_{F_{kr}}$ in (6) is estimated by the histogram of the frame $F_{kr}$, while the joint probability $p_{F_{kr}f_i}$ is estimated by the joint histogram of the frames $F_{kr}$ and $f_i$.

The closer the value of MI between these two frames is to zero, the less information one frame contains about the other. Therefore, the frame $F_{kr}$ is assigned to the pose class corresponding to the maximum $\mathrm{MI}(F_{kr},f_i)$. Because the pixel values of the images are used directly in Ref. 8, its estimation rate is very sensitive to illumination changes and noise. In order to obtain noise and illumination-invariant facial pose estimation algorithm, a new algorithm based on the feature face extracted by EMD and MI is proposed below.

First, we apply EMD to every partitioned video frame

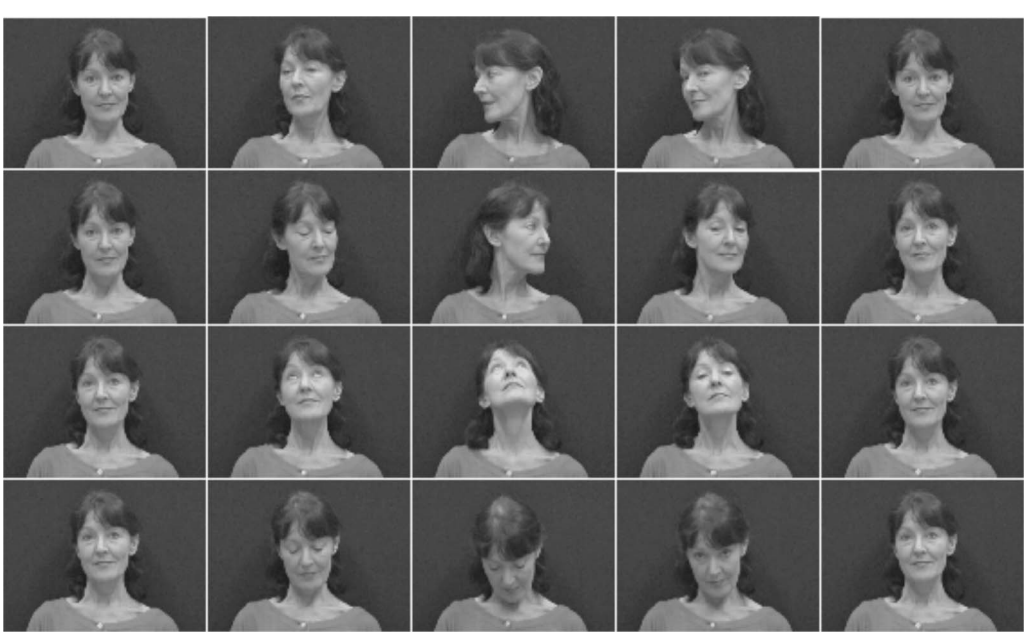

**Fig. 6** Samples of the VidTIMIT video sequence. The person starts from the frontal pose, turns her head to the left, right, back to the center, up, down, and then returns to center.

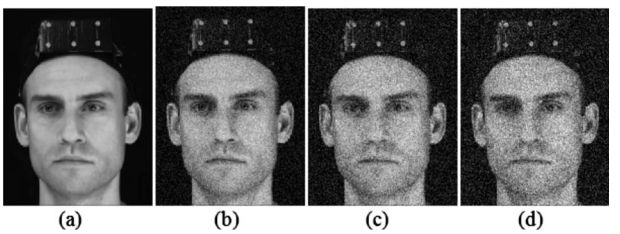


**Fig. 7** Examples of noise corrupted images: (a) the original image (b) corrupted image by variance $v=0.01$, (c) corrupted image by variance $v=0.02$, and (d) corrupted image by variance $v=0.03$.

$F_{kr}$ and the template image $f_i$. Consequently, the feature face $\tilde{F}_{kr}$ for the video frame $F_{kr}$ and the feature face $\tilde{f}_i$ for the template image $f_i$ can be obtained based on Eq. (4). By using MI, the pose estimation can be designed as such that we calculate the mean MI between $\tilde{F}_{kr}$ and $\tilde{f}_i$ as follows:

$$\overline{\mathrm{MI}}(F_{kr},O_{j,r})=\frac{1}{N(O_{j,r})}\sum_{f_i\in O_{j,r}}\mathrm{MI}(\tilde{F}_{kr},\tilde{f}_i), \tag{8}$$

where $f_i \in O_{j,r}$ and $N(O_{j,r})$ stands for the number of templates inside $O_{j,r}$. The MI of two feature faces, $\tilde{F}_{kr}$ and $\tilde{f}_i$, is defined as follows:

$$\mathrm{MI}(\tilde{F}_{kr},\tilde{f}_i)=H(\tilde{F}_{kr})+H(\tilde{f}_i)-H(\tilde{F}_{kr},\tilde{f}_i), \tag{9}$$

$$H(\tilde{F}_{kr})=-\sum_{u}p_{\tilde{F}_{kr}}(u)\log p_{\tilde{F}_{kr}}(u), \tag{10}$$

$$H(\tilde{F}_{kr},\tilde{f}_i)=-\sum_{u,v}p_{\tilde{F}_{kr}\tilde{f}_i}(u,v)\log p_{\tilde{F}_{kr}\tilde{f}_i}(u,v), \tag{11}$$

where $H(\tilde{F}_{kr})$ and $H(\tilde{f}_i)$ denote the marginal entropy values of $\tilde{F}_{kr}$ and $\tilde{f}_i$, respectively, and $H(\tilde{F}_{kr},\tilde{f}_i)$ is the joint entropy of the joint probability distribution of the image intensities. The probability $p_{\tilde{F}_{kr}}$ can be estimated by the histogram of the frame $\tilde{F}_{kr}$, while the joint probability density function $p_{\tilde{F}_{kr}\tilde{f}_i}$ can be estimated by a joint histogram of $\tilde{F}_{kr}$ and $\tilde{f}_i$.

**Table 1** Comparison on the MPI database [estimation rate (in percent)]

| | IMF1+MI | Residue+MI | The Proposed | MI[a] |
|---|---|---|---|---|
| Original images | 94.4 | 85.0 | 93.1 | 86.9 |
| Corrupted images $v=0.01$ | 31.8 | 79.8 | 87.3 | 85.4 |
| Corrupted images $v=0.02$ | 30.3 | 79.2 | 86.0 | 84.7 |
| Corrupted images $v=0.03$ | 29.5 | 79.4 | 85.4 | 84.6 |

[a]Reference 8.

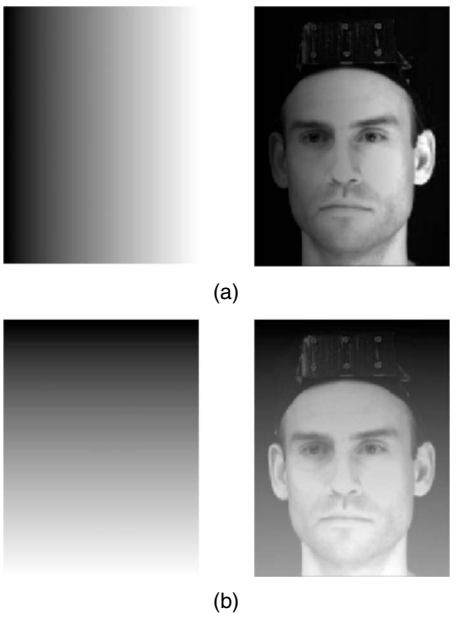


**Fig. 8** Examples of images with uneven lighting change. (a) Light from right to left and the corrupted image and (b) light from down to up and the corrupted image.

Therefore, the frame $F_{kr}$ is assigned to the pose class corresponding to the maximum $\overline{\mathrm{MI}}(F_{kr},O_{j,r})$, which is given as follows:

$$\mathrm{pose}=\arg\max_{j}\overline{\mathrm{MI}}(F_{kr},O_{j,r}). \tag{12}$$

## 4 Experimental Results and Discussions

In this section, extensive experiments are designed to evaluate the proposed facial pose-estimation algorithm in comparison to the existing MI based method in terms of their robustness to facial expressions, noise, and illumination variations. To make evaluations comprehensive, our experiments are performed on two publicly available face video databases:

1. Face Video Database of the Max Planck Institute (MPI) for Biological Cybernetics.[31,32] The cameras recorded 25 fps at 786×576 video resolution, noninterlaced. On the basis of the camera views, we have six different poses as shown in Fig. 4. From the video database, we extracted 38 action units for each pose, which illustrate different facial expressions. Ex-

**Table 2** Estimation rate (in percent) on the MPI database.

| | IMF1+MI | Residue+MI | Proposed | MI[a] |
|---|---|---|---|---|
| Original images | 94.4 | 85.0 | 93.1 | 86.9 |
| Light from right to left | 90.8 | 69.5 | 90.5 | 81.0 |
| Light from down to top | 94.4 | 53.6 | 93.1 | 66.1 |

[a]Reference 8.

**Table 3** Estimation rate (in percent) on the MPI database.

| | IMF1+MI | Residue+MI | Proposed | MI[a] |
|---|---|---|---|---|
| Original images | 94.4 | 85.0 | 93.1 | 86.9 |
| Noise+Light (right to left) | 29.6 | 66.8 | 84.8 | 76.2 |
| Noise+Light (down to top) | 31.7 | 51.7 | 87.1 | 57.3 |

[a]Reference 8.

amples are shown in Fig. 5. As a result, we have 228 video sequences (6 poses × 38 action units) in total to be processed.

2. VidTIMIT Database.[33] It is comprised of video and corresponding audio recordings of 43 people of three sessions recorded in different weeks. The content of each session is arranged as such that the subject performed a head-rotation sequence as shown in Fig. 6. The video of each person is stored as a numbered sequence of JPEG images with a resolution of 512 × 384 pixels. In total, there are 1806 (3 sessions × 9 poses × 2 tests to every session × 43 persons) tests to be carried out.

In our experimental framework, video is processed frame by frame. To achieve a better and more accurate verification rate, the algorithm resizes each video when the subject is in frontal position, according to a factor produced by a given standard distance between the right and left eyes using the method described in Ref. 34. In this way, the scaling problem occurring in different sessions of the same person can be resolved. From Ref. 8, we know that the best pose-estimation results are given for a specific dimension of the tracker bounding box where the face area is well described. In this paper, the size of the bounding box, in which the examined area only contains facial information, is 220 × 288 and 200 × 250, for the MPI and VidTIMIT databases, respectively. For each of the persons examined, a ground-truth image representing the required pose is used. This image is always taken from a different session from the one examined. The ground-truth constitutes $O_{j,r}$ in Eq. (8), while $F_{kr}$ is every following frame of the examined video input. In addition, the estimation rate on a database used in this paper is defined as

$$\text{Estimation Rate} = \frac{\text{No. frames(pose correctly estimated)}}{\text{No. frames in every video} \times \#\text{poses} \times \#\text{sessions} \times \#\text{tests to every session} \times \#\text{persons}}. \quad (13)$$

### 4.1 *Robustness to Noise*

In this section, experiments are designed to assess the robustness of the proposed algorithm. Considering the fact that the selected MPI database has 38 different facial expressions for each pose, we choose the normal facial expression as the ground-truth images for each pose, as shown in Fig. 4. Specifically, the video sequences are corrupted by Gaussian white noise with mean $m=0$ and variance set as $v=0.01$, 0.02, and 0.03. To quantify the difference between the corrupted image and the original image, the peak signal-to-nose ratio (PSNR) values are calculated, which is 30.11, 29.64, and 29.39 dB, for $v=0.01$, 0.02, and 0.03, respectively. Some samples of such noise corrupted images are shown in Fig. 7. We then applied the proposed algorithm to these noise-corrupted images to complete its pose estimation. For benchmarking purposes, we also tested the MI pose-estimation technique reported in Ref. 8, and all the results are summarized in Table 1.

It can be seen that when the video sequences in MPI database are corrupted by adding noise, the pose-stimation results are affected, which are decreased compared to the results tested on the original database, no matter which features we use. Especially, when the first IMF face is selected for pose estimation, the results become very poor with estimation rate of ~30%. However, when the residue faces are selected for pose estimation, the results did not change too much by varying levels of the added noise and the estimation rate is fixed at ~79%. Such huge difference confirms that the first IMF face contains the majority of the highest-frequency oscillations, which are essentially the variation of noises. In other words, without the local information of the first IMF, the proposed algorithm achieves a

**Table 4** Comparisons on the VidTIMIT database [estimation rate (in percent)]

| Method | Right | Mid-Right | Frontal | Mid-Left | Left | Up | Mid-Up | Mid-Down | Down |
|---|---|---|---|---|---|---|---|---|---|
| MI[a] | 90.0 | 70.0 | 99.8 | 68.3 | 83.3 | 85.0 | 90.0 | 55.0 | 95.0 |
| Proposed | 96.7 | 66.7 | 99.9 | 70.0 | 90.5 | 95.0 | 93.3 | 61.7 | 93.3 |

[a]Reference 8.

**Fig. 9** Successful pose estimation. The first row is the ground truth images, and the second row is the results produced.

certain level of robustness to noises. In addition, with the feature face proposed in this paper, better results are achieved than using the original intensities of facial image and the proposed algorithm outperforms the benchmark[8] in all tests. Although the proposed method is able to estimate 93.1% of the required poses on the original database, given the fact that there are large facial expression variations inside the database, the benchmark can only estimate 86.9% of the required poses.

### 4.2 *Robustness to Illumination*

The proposed algorithm has also been tested with respect to its sensitivity to illumination on the MPI database. Two typical uneven illumination variations are added to the database as shown in Fig. 8. The simulation results are illustrated in Table 2. It can be seen that the results of the first IMF face and the proposed method are not affected by the illumination changes. However, the results of the residue and the method in Ref. 8 are very sensitive to lighting changes and their estimation rates are deteriorated. This illustrates that the decomposed residue by EMD contains the majority of illumination variations. For vertical light changes, the pixel values of each row in the corrupted image presents the effect of being added a constant value compared to the original image. Following the EMD decomposition, such effect will not have any impact on the IMFs but the residues. In other words, the IMFs of the original image will remain the same as that of the corrupted image and the difference only exists in the residues, which often present a constant value. Therefore, under such circumstance of corruptions, the proposed algorithm can still maintain the 93.1% estimation rate, while the benchmark in Ref. 8 can only achieve an estimation rate of 66.1%.

In practical applications, a facial image is often contaminated by noise and uneven illumination variations at the same time. Table 3 illustrates the simulation results on the MPI database when images are corrupted by Gaussian white noise with mean $m=0$ and variance $v=0.01$, and inhomogeneous illumination changes from right to left or from down to top. It can be seen that the proposed algorithm can still find correctly 84.8 or 87.1% poses when images are contaminated by noise and illumination together. However, under such circumstance of corruptions, the benchmark in Ref. 8 can only achieve an estimation rate of 76.2 or 57.3%.

### 4.3 *Comparisons on the VidTIMIT Database*

To further evaluate the proposed algorithm in comparison to existing efforts, we used the VidTIMIT database to test the proposed algorithm on nine poses with different video sessions, while the algorithm reported in Ref. 8 only tested seven poses. The results of the comparison on each pose are given in Table 4. It can be seen that both of these methods can achieve the highest estimation rate when dealing with the frontal pose. However, the lowest estimation rate of the mid-down pose for MI in Ref. 8 is 55.0%, whereas it is 61.7% for the proposed. Furthermore, the average estimation rate of all the poses is 81.82% for MI in Ref. 8 and 85.23% for the proposed. Therefore, the proposed algorithm achieves better performances for almost all poses and some examples of successful estimation are presented in Fig. 9.

## 5 Conclusions

In this paper, a new facial pose-estimation algorithm is proposed based on the EMD and MI. In order to be invariant to noise and illumination, a new feature face is designed, which is reconstructed by using IMFs that do not contain noise and illumination effects. In comparison to existing work, the proposed algorithm achieves the following advantages: (*i*) utilizes EMD as a bandpass filter to extract discriminating feature faces for facial pose estimation; (*ii*) being robust to Gaussian white noise corruption, uneven illumination variation and varying facial expressions; and (*iii*) outperforms the existing MI-based pose-estimation algorithm[8] with more poses tested for evaluation in terms of both pose-estimation accuracy and pose-estimation variations. Experimental results on both the MPI and the VidTIMIT video databases validate the advantages of the proposed algorithm and show that EMD is proved to be a powerful tool for feature extraction and its application to pose estimation provides excellent potential for further multimedia content analysis and pattern recognitions

### *Acknowledgments*

The authors acknowledge the funding support from European Framework-7 Research Programme under the HERMES project (Contract No. 216709), Natural Science Foundation of Guangdong Province of China (Grant No. 9451027501002552), and National Natural Science Foundation of China (Grant No. 10926139).

## References


1. E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(4), 607–626 (2009).
2. J. Sherrah, S. Gong, and E. Ong, "Face distributions in similarity space under varying head pose," *Image Vis. Comput.* **19**(12), 807–819 (2001).
3. Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 300–305 (2000).
4. Y. Wei, L. Fradet, and T. Tan, "Head pose estimation using Gabor eigenspace modeling," in *Proc. of the IEEE Int. Conf. on Image Processing*, Vol. **1**, pp. 281–284 (2002).
5. L. Chen, L. Zhang, Y. Hu, M. Li, and H. Zhang, "Head pose estimation using fisher manifold learning," in *Proc. of the IEEE Int. Workshop on Analysis and Modeling of Faces and Gestures*, in conjunction with ICCV2003, pp. 203–207 (2003).
6. S. Z. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng, "Learning multiview face subspaces and facial pose estimation using independent component analysis," *IEEE Trans. Image Process.* **14**(6), 705–712 (2005).
7. J. Wu and M. M. Trivedi, "A two-stage head pose estimation framework and evaluation," *Pattern Recogn.* **41**, 1138–1158 (2008).
8. G. Goudelis, A. Tefas, and I. Pitas, "Automated facial pose extraction from video sequences based on mutual information," *IEEE Trans. Circuits Syst. Video Technol.* **18**(3), 418–424 (2008).
9. S. S. Wang, P. C. Chen, and W. G. Lin, "Invariant pattern recognition by moment Fourier descriptor," *Pattern Recogn.* **27**, 1735–1742 (1994).
10. P. Wunsch and A. F. Laine, "Wavelet descriptors for multiresolution recognition of handprinted characters," *Pattern Recogn.* **28**, 1237–1249 (1995).
11. S. Mallat, *A wavelet tour of signal processing*, 2nd ed., Elsevier, New Yark (1999).
12. N. E. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and nonstationary time series analysis," *Proc. R. Soc. London* **454**, 903–995 (1998).
13. W. Huang, Z. Shen, N. Huang, and Y. Fung, "Engineering analysis of biological variables: an example of blood pressure over 1 day," in *Proc. Nat. Acad Sci USA* **95**, 4816–4821 (1998).
14. H. Liang, Z. Lin, and R. W. McCallum, "Artifact reduction in electrogastrogram based on empirical mode decomposition method," *Med. Biol. Eng. Comput.* **38**, 35–41 (2000).
15. Z. Yang, L. Yang, D. Qi, and C. Y. Suen, "An EMD-based recognition method for Chinese fonts and styles," *Pattern Recogn.* **27**, 1692–1701 (2006).
16. N. Bi, Q. Sun, D. Huang, Z. Yang, and J. Huang, "Robust image watermarking based on multiband wavelets and empirical mode decomposition," *IEEE Trans. Image Process.* **16**(8), 1956–1966 (2007).
17. R. Bhagavatula and M. Savvides, "Analyzing facial images using empirical mode decomposition for illumination artifact removal and improved face recognition," presented at IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2007, ICASSP 2007, Vol. **1**, Issue 15–20, pp. I-505–I-508 (2007).
18. G. Rilling, P. Flandrin, and P. Gongalves, "On empirical mode decomposition and its algorithms," presented at IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (I) (2003).
19. P. Flandrin, G. Rilling, and P. Gonçalvés, "Empirical mode decomposition as a filter bank," *IEEE Signal Process. Lett.* **11**(2), 112–114 (2004).
20. B. B. Mandelbrot and J. W. Van Ness, "Fractional brownian motions, fractional noises and applications," *SIAM Rev.* **10**, 422–437 (1968).
21. P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data," in *Self-Similar Network Traffic and Performance Evaluation*, K. Park and W. Willinger, Eds., pp. 39–88, Wiley, Hoboken, NJ (2000).
22. Z. Wu and N. E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method," *Proc. R. Soc. London, Ser. A* **460**(2046), 1597–1611 (2004).
23. C. P. Chang, J. C. Lee, Y. Su, P. S. Huang, and T. M. Tu, "Using empirical mode decomposition for iris recognition," *Comput. Standards Interfaces* **31**(4), 729–739 (2009).
24. J. C. Nunes, S. Guyot, and E. Delechelle, "Texture analysis based on local analysis of the bi-dimensional empirical mode decomposition," *Mach. Vision Appl.* **16**, 177–188 (2005).
25. C. Damerval, S. Meignen, and V. Perrier, "A fast algorithm for bidimensional EMD," *IEEE Signal Process. Lett.* **12**(10), 701–704 (2005).
26. Y. Xu, B. Liu, and S. Riemenschneider, "Two-dimensional empirical mode decomposition by finite elements," *Proc. R. Soc. London, Ser. A* **462**, 3081–3096 (2006).
27. G. Xu, X. Wang, and X. Xu, "Improved bi-dimensional EMD and Hilbert spectrum for the analysis of textures," *Pattern Recogn.* **42**, 718–734 (2009).
28. N. E. Huang and Z. Wu, "A review on Hilbert-Huang transform: method and its applications to geophysical studies," *Rev. Geophys.* **46**, RG2006 (2008).
29. P. Viola and W. M. Wells, III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.* **24**, 137–154 (1997).
30. F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imaging* **16**, 187–198 (1997).
31. C. Wallraven, D. W. Cunningham, M. Breidt, and H. H. Bulthoff, "View dependence of complex versus simple facial motions," in *Proc. of 1st Symp. on Applied Perception in Graphics and Visualization* p. 181, ACM Press, New York (2004).
32. M. Kleiner, C. Wallraven, and H. H. Bulthoff, "The MPI VideoLab—a system for high quality synchronous recording of video and audio from multiple viewpoints," MPI-Tech. Reports, No. 123, (2004).
33. C. Sanderson, *Biometric person recognition: face, speech and fusion*, VDM-Verlag, Germany (2008).
34. S. Asteriadis, N. Nikolaidis, I. Pitas, and M. Pardas, "Detection of facial characteristics based on edge information," in *Proc. of 2nd Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, pp. 247–252, Barcelona, Spain (2007).



**Chunmei Qing** received her BS in mathematics and computing science from Sun Yat-sen University, China, in 2003, and PhD in electronic imaging and media communications from University of Bradford, United Kingdom, in 2009. Her current research interests include image/video processing, time-frequency analysis, and pattern recognition.

**Jianmin Jiang** received his BSc from Shandong Mining Institute, China, in 1982, MSc from the China University of Mining and Technology in 1984, and PhD from the University of Nottingham, Nottingham, United Kingdom, in 1994. From 1985 to 1989, he was a lecturer with Jiangxi University of Technology, China. In 1989, he joined Loughborough University, United Kingdom, as a visiting scholar and later moved to the University of Nottingham as a research assistant. In 1992, he was appointed a lecturer of electronics at the Bolton Institute, United Kingdom, and rejoined Loughborough University in 1995 as a lecturer of computer science. In 1997, he was appointed as a full professor at the School of Computing, University of Glamorgan, Pontypridd, United Kingdom, and joined the University of Bradford, United Kingdom, in 2002 as the chair of digital media at the School of Informatics. His research interests include visual information retrieval, image/video processing, visual content management, Internet video coding, stereo image coding, and neural network applications. He has authored more than 150 refereed published research papers. Dr. Jiang is a Fellow of the Institute of Electrical Engineers (IEE) and the RSA U.K.

**Zhijing Yang** received his BS and PhD in mathematics and computing science from Sun Yat-sen University, China, in 2003 and 2008, respectively, where he is currently a lecturer in the School of Mathematics and Computing Science. His research interests include signal processing, time-frequency analysis, and pattern recognition.